Change detection in hydrological records— a review of the methodology

ZBIGNIEW W. KUNDZEWICZ

Research Centre of Agricultural and Forest Environment, Polish Academy of Sciences, Bukowska 19, 60-809 Poznań, Poland, and
Potsdam Institute for Climate Impact Research, D-14412 Potsdam, Germany
zkundze@man.poznan.pl

ALICE J. ROBSON

Centre for Ecology and Hydrology, Wallingford, Oxfordshire OX10 8BB, UK

Abstract General guidance is offered as to the methodology of change detection in time series of hydrological data, embracing stages such as preparing a suitable data set, exploratory analysis, application of adequate statistical tests and interpretation of results. Although the paper cannot go into full details of the many existing tests, it gives an easy-to-follow overview, offering practical hints and describing caveats and misconceptions. It serves as a refresher, raising attention to essential things that have often been ignored. A particular recommendation of the paper is that greater use of distribution-free testing methods, particularly resampling methods, should be made. These methods are recommended because they are particularly suited to hydrological data, which are often strongly skewed (non-normal), seasonal and serially correlated. Resampling techniques are flexible, robust and powerful, and require only minimal assumptions to be made about the data.

Key words change detection; trend; statistical testing; resampling; hydrological data

Revue méthodologique de la détection de changements dans les chroniques hydrologiques

Résumé Nous offrons un guide général des méthodes de détection de changement dans les séries chronologiques de données hydrologiques, comprenant les étapes de préparation d'un jeu de données convenable, d'analyse exploratoire, d'application des tests statistiques appropriés et d'interprétation des résultats. Bien que l'article ne puisse pas présenter en détail tous les tests existants, il donne une vision générale facile à suivre, offre des conseils pratiques, exprime quelques mises en garde et dénonce quelques idées fausses. Il sert de rappel, en attirant l'attention sur des points essentiels souvent ignorés. Une recommandation de l'article est que les méthodes basées sur des tests non-paramétriques, en particulier les méthodes de rééchantillonage, devraient être plus utilisées. Ces méthodes sont recommandées parce qu'elles conviennent particulièrement aux données hydrologiques, qui sont souvent fortement asymétriques (non normales), saisonnières et autocorrélées. Les techniques de rééchantillonnage sont flexibles, robustes et puissantes, et requièrent seulement des hypothèses minimales par rapport aux données.

Mots clefs détection de changement; tendance; test statistique; rééchantillonnage; données hydrologiques

INTRODUCTION

Detection of changes in long time series of hydrological data is an important and difficult issue, of increasing interest. The aim of the present paper is to summarize the essential components of a change detection study, and to offer broad recommendations and general guidance for methodology that is particularly suited to change detection in typical hydrological data.

This paper partly draws from a report of a WMO/UNESCO/CEH workshop set up to review trend detection in hydrological series (Kundzewicz & Robson, 2000). The workshop, held at the Centre for Ecology and Hydrology in Wallingford, UK, in December 1998, brought together international experts from the hydrological and statistical community, and sought to make general recommendations on methodology for use in detection of trend and other change in hydrological data.

This paper attempts to remind investigators of the many necessary aspects that should form part of a trend detection study and discusses the key components of an analysis. The paper considers preparation of a suitable data set, emphasizing the importance of fully understanding the data. It briefly discusses exploratory analysis of the data, a visual approach that should form an integral part of any study of change. It also considers interpretation of the results, and the need for examining as much external evidence as possible in order to distinguish between trend that arises from land-use change, climate change, or simply from altered measurement techniques.

The main component of the paper looks at methods for hydrological change detection. It introduces distribution-free methods and resampling methods that are particularly suited to hydrological data and can be used in a wide range of situations under minimal data assumptions.

Although the present contribution cannot go into full details of the many existing tests, it gives an easy-to-follow overview, offering practical hints, and describing caveats and misconceptions. It serves as a refresher, raising attention to essential things that have often been ignored.

DATA FOR HYDROLOGICAL CHANGE DETECTION

Data are the backbone of any attempt to detect trend or other change in hydrological records. It is easy to neglect the importance of properly preparing and understanding the data to be used, and the necessity of working with data of as high a quality as possible. Data should be quality-controlled before commencing an analysis of change. Examples of problems linked to the data that can cause apparent changes in a data series are: typographical errors; instrument malfunction (zero-drift, bias); and change in measurement techniques, in instrumentation, or in instrument location, in accuracy of data, or in data conversions (e.g. altered rating equations). Such changes may occur either through time or at different sites, and these need to be identified. One should be open-minded at any stage of an analysis and keep an eye out for possible quality problems that may previously have been missed. Missing values and gaps in data are complicating factors. Whether or not to fill them, and if so, in what way, is a complex issue. The problem needs particular care when the gaps are non-random, e.g. following equipment damage from a flood event of exceptional magnitude.

It is important to consider carefully the form and frequency of the data that should be analysed. This usually depends on the focus of the study. The form in which the data are collected is not always the most appropriate for the study in question. For floods, the biggest flow is often of interest; for droughts, it may be the duration of low flows. Processing very frequent data is computationally intensive and such data may be highly correlated in time. It can be worth simplifying the data by reducing the frequency (aggregating) or using summary measures, such as maxima or averages. Use of data transformation (e.g. taking logarithms, or using ranks and normal scores) can compensate for undesirable data properties, e.g. high skewness or strong departure from normality.

Selection of which stations to use in a study is also important. For example, the issue of detecting a climate change signature in river flow data is very complex because the process of river flow is the integrated result of several factors, such as precipitation inputs, catchment storage and evaporation losses. Furthermore, climate change signals may be overshadowed by strong natural background variability. These factors mean that particular care is needed in selecting data and sites for use in studying climate change. In order to study climate change signature in river flow, data should ideally be taken from pristine/baseline rivers and should be of high quality and extend over a long period. Where pristine sites are not available, it may be possible to eliminate other influences, or to conceptually reconstruct natural flows. Detailed suggestions on how to select a network of stations for climate change detection are given in Pilon (2000).

EXPLORATORY DATA ANALYSIS

Exploratory data analysis (EDA) is an advanced visual examination of the data. It involves using graphs to explore, understand and present data, and is an essential component of any statistical analysis. A study of change that does not include a thorough exploratory data analysis is not complete. The first use of EDA is usually to examine the raw data in order to identify such features as data problems (outliers, gaps in the record, etc.); temporal patterns (e.g. trend or step-change, seasonality); and regional and spatial patterns. Exploratory data analysis also plays an important role in checking out test assumptions such as independence, or statistical distribution of data values. Finally, EDA is invaluable when it comes to interpreting and presenting the results of a statistical analysis, e.g. for examining residuals, trend gradients and significance levels.

Exploratory data analysis allows a much greater appreciation of the features in data than tables of summary statistics and statistical significance levels. This is because the human brain and visual system is very powerful at identifying and interpreting patterns. It is often able to see important features, structures or anomalies in a data series that would be very difficult to detect in any other way. Just looking at the data can change initial preconceptions, can alter the questions that it is sensible to ask, and can uncover important aspects that might not otherwise have been found.

A well-conducted EDA is such a powerful tool that it can sometimes eliminate the need for a formal statistical analysis. Alongside EDA, statistical tests become a way of confirming whether an observed pattern is significant, rather than a means of searching through data. Further, EDA is often useful in identifying data quality problems. However, it is not a substitute for proper quality control of data.

A good EDA involves plotting, studying and refining graphs so as to highlight important features of the data and thus identify further graphs that are needed. Deciding which graphs to look at is a matter of judgement and experimentation. Common types of graph that can be useful for hydrological data series include histograms and normal probability plots, time series plots, autocorrelation plots, scatter plots and smoothing curves. For further details and examples applicable to hydrological data, the user should refer to Grubb & Robson (2000). Excellent presentations of the general principles of EDA can also be found in Cleveland (1994) and Tufte (1983).

APPLICATION OF STATISTICAL TESTS

Basics of statistical testing for change

Change in a series can occur in numerous ways: e.g. gradually (a *trend*), abruptly (a *step-change*), or in a more complex form. It may affect the mean, median, variance, autocorrelation, or almost any other aspect of the data.

In order to carry out a statistical test, it is necessary to define the null and alternative hypotheses; these are statements that describe what the test is investigating. For example, to test for trend in the mean of a series the null hypothesis (H_0) would be that there is no change in the mean of a series, and the alternative hypothesis (H_1) would be that the mean is either increasing or decreasing over time. In carrying out a statistical test, one starts by assuming that the null hypothesis is true, and then checks whether the observed data are consistent with this hypothesis. The null hypothesis is rejected if the data are not consistent.

To compare between the null and alternative hypotheses, a test statistic is selected and then its significance is evaluated, based on the available evidence. The test statistic is simply a numerical value that is calculated from the data series that is being tested, which is selected so that it highlights the difference between the two hypotheses. A simple example of a test statistic is the linear regression gradient; this can be used to test for a trend in the mean. If there is no trend (the null hypothesis) then the regression gradient should have a value near to zero. If there is a large trend in the mean (the alternative hypothesis), then the value of the regression gradient would be very different from zero, being positive for increasing trend and negative for decreasing trend.

The significance level measures whether the test statistic is very different from the range of values that would typically occur under the null hypothesis. It is the probability that a test erroneously detects trend when none is present. Thus a 5% significance level would be interpreted as strong evidence against the null hypothesis—with a 1 in 20 chance of that conclusion being wrong. The significance level expresses the probability that the null hypothesis is incorrectly rejected; incorrect rejection of the null hypothesis is known as type I error (Table 1).

Another type of error (type II error, see Table 1) occurs when the null hypothesis is accepted (i.e. no trend is present) when the alternative hypothesis is true (a trend

Table 1 Interpretation of errors of type I and II.

		Does a trend exist?	
		Yes	**No**
Has a trend been detected?	**Yes**	+	Error of type I: false trend detected when none exists
	No	Error of type II: failure to detect an existing trend (e.g. due to weakness of the trend, or of the methodology, or shortness of the record)	+

exists). When a test has low type II error probability (i.e. the risk of incorrectly accepting the null hypothesis is low), the test is said to be powerful, and more powerful tests are to be preferred. The power of the test is the probability of correctly detecting a trend when one is present.

In carrying out a statistical test it is always necessary to consider assumptions, such as:

A specified form of distribution (e.g. assuming that the data are normally distributed). This assumption is violated if the data do not follow the specified distribution.

Constancy of the distribution (i.e. all data points have an identical distribution) This assumption is violated if there are seasonal variations or any other cycles in the data, or if there is an alteration over time in the variance or any other feature of the data that is not allowed for in the test.

Independence This assumption is violated if there is autocorrelation (correlation from one time value to the next; also referred to as serial correlation or temporal correlation) or, in the case of a multi-site study, spatial correlation (correlation between sites).

If the assumptions made in a statistical test are not fulfilled by the data, then test results can be meaningless, in the sense that the estimates of significance level would be grossly incorrect. Hydrological data are often strongly non-normal, and this means that tests which assume an underlying normal distribution are not adequate. Hydrological data also typically either show autocorrelation and/or spatial correlation and, therefore, data values are not independent. They may also display seasonality, which violates assumptions of constancy of distribution.

An outline of the main stages of a statistical analysis of change

The main stages in statistical testing are:

– Decide what type of series/variable to test depending on the issues of interest (e.g. monthly averages, annual maxima, deseasonalized data, etc.).
– Decide what types of change are of interest (gradual trend or step-change).
– Check out data assumptions (e.g. use exploratory data analysis, or a formal test).
– Select a statistical test (more than one is good practice). This means selecting a test statistic and selecting a method for evaluating significance levels.
– Evaluate significance levels.
– Investigate and interpret results.

Distribution-free testing

There are many ways of testing for trend or other change in hydrological data. These range from traditional approaches such as maximum likelihood estimation, and Bayesian and time series methods, to newer approaches such as phase randomization

(Radziejewski *et al.*, 2000) and smoothing techniques (e.g. locally weighted regression). Each of these methods has a time and a place for use.

In this paper, the focus will be on a particular group of methods called distribution-free methods, and in particular, on one group of distribution-free methods known as resampling. A distribution-free method is one that does not require any assumption about the form of distribution that the data derive from, e.g. there is no need to assume data are normally distributed.

The following approaches are distribution-free:

Rank-based tests These use the ranks of the data values (not the actual data values). A data point has rank r if it is the rth largest value in a data set. There are a number of widely used and useful rank-based tests. Most rank-based tests assume that data are independent and identically distributed. Rank-based tests have the advantage that they are robust and usually simple to use. They are usually less powerful than a parametric approach.

Tests using a normal scores transformation There are many tests for change that rely on assumptions of normality. Such tests are generally not suitable for direct use with hydrological data, which are, typically, far from being normally distributed. However, such tests can be used if the data are first transformed. The normal scores transformation results in a data set that has a normal distribution. It is similar to using the ranks of a data series, but instead of replacing the data value by its rank, r, the data value is replaced by the typical value that the rth largest value from a sample of normal data would have (the rth normal score). The advantages of using normal scores are that the original data need not follow a normal distribution, and the test is relatively robust to extreme values. The disadvantage is that statistics measuring change, such as the regression gradient, cannot be easily interpreted. Normal scores tests are likely to give slightly improved power for detection of change relative to equivalent rank-based tests.

Tests using resampling approaches Resampling methods use the data to determine the significance of a test statistic. They are described in more detail below.

Introducing resampling methods

Resampling methods (permutation or bootstrap) are robust techniques for estimating the significance level of a test statistic. A useful practical text on resampling and permutation is provided by Good (1993), while Efron & Tibshirani (1998) and Davidson & Hinkley (1997) describe bootstrapping methods. The advantages of resampling are that it is a flexible method that can be adapted to a wide range of types of data, including autocorrelated or seasonal data. Resampling is also relatively powerful, e.g. for large samples, permutation tests can be shown to be as powerful as the most powerful parametric tests (Bickel & Van Zwet, 1978). Resampling methods are very useful for testing hydrological data because they require relatively few assumptions to be made about the data, yet they are also quite powerful tests.

Understanding resampling

The basic idea behind resampling methods is very straightforward. Consider testing a series for trend: a possible test is the regression gradient. If there is no trend in the data (the null hypothesis), then the order of the data values should make little difference. Thus shuffling (permuting) the elements of the data series should not change the gradient very much. Under a permutation approach, the data are shuffled very many times. After each shuffle (permutation) the test statistic is recalculated. After very many permutations, the original test statistic is compared to the generated test statistic values. If the original test statistic is rather different from most of the generated values, then this suggests that the ordering of the data affects the gradient and thus that there was trend. If the original test statistic lies somewhere in the middle of the generated values, then it seems reasonable that the null hypothesis was correct (the order of the values does not matter, so there is no evidence of trend). In other words, if an observer (or in this case, the statistical test) can distinguish between the original data and the resampled (permuted) data, then the observed data are judged not to satisfy the null hypothesis.

Permutation and the bootstrap

The bootstrap and permutation methods are two slightly different approaches to resampling the data. In permutation methods (sampling with no replacement) the data are re-ordered, i.e. each of the data points in the original data series appears only once in each resampled (generated) data series. In bootstrap methods, the original data series is sampled with replacement to give a new series with the same number of values as the original data. The series generated with this method may contain more than one of some values from the original series and none of other values. In both cases, the generated series has the same distribution as the empirical (i.e. observed) distribution of the data.

The bootstrap is generally, but not always, less powerful than a permutation test (Good, 1993). However, bootstrap methods are often preferred where a test is looking for change in variance. Further, permutation tests cannot be applied with test statistics that do not change when the data are permuted, e.g. tests for which the test statistic is the mean or median. In general, bootstrap methods are more flexible than permutation methods and can be used in a wider range of circumstances.

Number of resamples

The number of resamples depends on the level of significance required and on the degree of change in the data. The larger the number of resamples, the more accurate the estimate of significance. More resamples will be required to accurately determine significance levels of 1% than of 10%. A simple approach to check whether the sample size is sufficient is to re-run a test a few times and check that the required percentiles of the generated test statistic values do not vary too much.

For permutation testing, all permutations could, theoretically, be evaluated. However, typically there are too many to be evaluated (for a series of length n there are $n!$ permutations) and a random selection of possible permutations is used instead.

Note that if confidence intervals are required, then sample sizes of 199, 999, 1999, etc. give exact confidence intervals (Faulkner & Jones, 1999), e.g. for 199 samples, the 95% confidence interval is given by the fifth largest and smallest values; for 1999 samples, the 95% confidence interval is given by the 50th largest and smallest values.

Choice of resampling strategy

The simplest resampling strategy is to permute or bootstrap individual data points, as described above. This technique is applicable only in the case where it can be assumed that the data are independent and non-seasonal.

If data show autocorrelation, or additional structure such as seasonality, then it is necessary that the series generated by resampling should replicate this structure. A straightforward means of achieving this is to permute, or bootstrap the data in blocks. For example, for a 40-year series of monthly values, it would be sensible to treat the data as consisting of 40 blocks of one year. Each year's worth of data is left intact and is moved around together as a block, thus maintaining the seasonal and temporal dependencies within each year. The 40 blocks are then re-ordered many times. In this way, the resampled series would preserve the original seasonality. Similarly, blocks can be forced to replicate the autocorrelation in the data. It is important that the size of the blocks should be sensibly selected. If there is seasonality, then the block should contain an integral number of seasonal cycles. If there is autocorrelation, then the block should be chosen so that data points one block apart are approximately independent.

Note also that blocking methods can be useful when there is spatial dependency in a set of multi-site data that is to be tested as a group. In this case, the usual choice of blocks would be to group data across all sites that have data that occur in the same time interval (e.g. Robson *et al.*, 1998).

SOME COMMONLY USED TESTS AND TEST STATISTICS

This section lists a number of standard tests for detection of step-change and gradual trend. The tests are described in their standard or basic form, i.e. in a non-resampling framework. Each of these tests can be easily adapted to be a resampling test. For this, the same test statistic is calculated as for the basic test, but the significance level is

Table 2 General applicability of tests.

Situation	Guidelines for test selection
Data are normally distributed, independent and non-seasonal	This is an unlikely scenario for hydrological data. If applicable, any of the tests listed below should be suitable.
Data are independent and non-seasonal, but are non-normal	Any of the distribution-free tests are suitable. Tests that are based on normality assumption can also be applied, either by (a) first applying a normal scores or ranks transformation, or by (b) using a relevant test statistic and evaluating significance using resampling techniques.
Data are non-normal and are not independent, or are seasonal	For almost all the tests listed below, it will be necessary to extract the test statistic, and then to evaluate significance levels using block-permutation or block-bootstrap methods. Without this test assumptions will not be met. The exception is the Seasonal Kendall test, which may be used with seasonal data.

obtained using the resampling approach. Whether it is appropriate to use the basic test procedure (i.e. without resampling), will depend on the assumptions that can be made about the data (cf. Table 2).

Note that all tests assume that, under the null hypothesis, the distribution of data values does not change with either time or space. If this is not appropriate, then more sophisticated testing approaches will be necessary.

For resampling techniques, it is possible to construct new test statistics to test for a particular type of change; it is not necessary to select test statistics from known tests. Having the flexibility to construct "custom" test statistics allows great flexibility in what can be tested for.

Tests for step change

1. Median change point test/Pettitt's test for change This is a rank-based test for a change in the median of a series with the exact time of change unknown (Pettitt, 1979; Siegel & Castellan, 1988). The test is considered to be robust to changes in distributional form and relatively powerful, e.g. compared to the Wilcoxon-Mann-Whitney test (see below).

2. Wilcoxon-Mann-Whitney test/Mann-Whitney test/Mann-test/Rank-sum test This is a rank-based test that looks for differences between two independent sample groups (Siegel & Castellan, 1988; WMO, 1988; Helsel & Hirsch, 1992). It is based on the Mann-Kendall test statistic (see test 9 below), but is calculated for subsets of the series in order to detect the point of change in the mean (Chiew & McMahon, 1993). In its basic form it assumes that the time of change is known. When the time of change is unknown, use of the median change-point test is recommended.

3. Distribution-free CUSUM test This is a rank-based test in which successive observations are compared with the median of the series (Chiew & McMahon, 1993; McGilchrist & Woodyer, 1975). The test statistic is the maximum cumulative sum (CUSUM) of the signs of the difference from the median (i.e. the CUSUM of a series of values of +1 or –1) starting from the beginning of the series.

4. The Kruskal-Wallis test The Kruskal-Wallis test (Sneyers, 1975), is a rank-based test for equality of sub-period means. It can also be used to test for equality of sub-period variability.

5. Cumulative deviations and other CUSUM tests The cumulative deviation test (Buishand, 1982) is based on the rescaled cumulative sum of the deviations from the mean. The test is relatively powerful in comparison with other tests (e.g. Worsley likelihood ratio test; Buishand, 1982) for a change-point that occurs towards the centre of the time series. The basic test assumes normally distributed data. Other CUSUM based tests (using Bayesian and likelihood methods) are described in Buishand (1984).

6. Student's *t* test This is a standard parametric test for testing whether two samples have different means. In its basic form it assumes normally distributed data and a known change-point time.

7. The Worsley likelihood ratio test The Worsley likelihood ratio test (Worsley, 1979) is similar to Student's *t* test but is suitable for use when the change-point time is unknown. It assumes normality.

Tests for trend

8. Spearman's rho This is a rank-based test for correlation between two variables that can be used to test for a correlation between time and the data series (Siegel & Castellan, 1988). Spearman's correlation is a rank-based version of the usual parametric measure of correlation (the Pearson product moment; Sprent, 1989).

9. Kendall's tau/Mann-Kendall test This is another rank-based test, which is similar to Spearman's rho (same power and still based on ranks), but using a different measure of correlation which has no parametric analogue.

10. Seasonal Kendall test The Seasonal Kendall test is a version of the Mann-Kendall test that allows for seasonality in the data (Hirsch *et al.*, 1982). There is also a modified seasonal Kendall test that additionally allows for some autocorrelation in the data (Hirsch & Slack, 1984).

11. Linear regression The test statistic for linear regression is the regression gradient. This is one of the most common tests for trend and, in its basic form, assumes that data are normally distributed.

12. Other robust regression tests There are a number of robust methods for estimating trend in series. These could potentially be used as alternative measures of the change. For example, in least absolute deviation regression, the gradient is that which minimizes the sums of the deviations of the points from the fitted line (Bloomfield & Steiger, 1983). Other robust means of estimating the rate of change include M-estimates of regression and trimmed regression (Rousseeuw & Leroy, 1987).

INTERPRETING TEST RESULTS

When interpreting test results it is necessary to remember that no statistical test is perfect, even if all test assumptions are met. Assuming a 5% significance level means that an error will be made, on average, for 5% of the time: i.e. if the null hypothesis was true then one in 20 test results will be significant and incorrect. If more than one test has been applied to the data, interpretation of results can be complex. The presence of a single significant test result may only be weak evidence of change—even if this test is highly significant. If more tests are significant then this provides stronger evidence of change, unless they are very similar, in which case multiple significance is not an extra proof of change.

It is important to examine the test results alongside graphs of the data, and with as much historical knowledge about the data as possible. For example, if both step-change and trend results are significant, extra information will be needed to determine which of these provides the best description of the change. If historical investigations reveal that a dam was built during the period, and this is consistent with the time series plot, then a reasonable conclusion would be that the dam caused a step change. It can often be helpful to look out for patterns in the results that may indicate further structure, e.g. regional patterns in trends. These may suggest that further investigation is needed.

If a test indicates a significant change in a data series, then it is important to try to understand the cause. For example, the investigator may be interested in detecting climate change, but there may be many other possible explanations, such as changes caused directly by man (urbanization, reservoirs, drainage systems, water abstraction, land-use change etc.), natural catchment changes (e.g. natural changes in channel morphology), climate variability and problems linked to data (see earlier section).

It is very important to understand the difference between climate variability and change, where the former is the natural variation in the climate from one period to the next, while the latter refers to a long-term alteration in the climate. Climate variability appears to have a very marked effect on many hydrological series. This has two important consequences:

Climate variability can cause apparent trend Climate variability can easily give rise to apparent trend when records are short—these are trends that would be expected to disappear once more data had been collected. Because of climate variability, records of 30 years or less are almost certainly too short; at least 50 years of record is necessary for climate change detection.

Climate variability obscures other changes Because climate variability is typically large, it can effectively obscure any underlying changes due either to climate change or to urbanization.

The best way to improve understanding of change is to gather as much information as possible, using, e.g. information about changes in the catchment (land-use change, etc.) and about data collection methods. Data from nearby sites can also be useful—if they show similar patterns, then the cause is probably widespread (e.g. linked to climate, or to extensive land-use change). In addition, related variables (proxies) can be used—information on temperature and rainfall can help determine whether changes in flow can be explained by climatic factors. If related data can be obtained that extend to a longer period than the original data, this may be of assistance.

CONCLUDING REMARKS

This paper has outlined the key components required for a study of change in hydrological data. It has emphasized the critical role of accurate and well-understood data, of using exploratory data analysis as a key part of the analysis, of paying adequate attention to the validity of assumptions, of selecting tests and determining the significance levels, and of taking a wide perspective when interpreting test results.

This paper focuses on the use of distribution-free methods and, particularly, resampling methods for testing hydrological data. It briefly presents a variety of common tests. Distribution-free methods are recommended because they allow minimal assumptions to be made about the data and are therefore particularly suited to hydrological series, which are often neither normally distributed nor independent. Resampling techniques are a particularly flexible approach as they can be used even when data are autocorrelated or seasonal, by employing block-permutation or block-bootstrap techniques.

The present contribution provides general guidance and recommendations for detecting change in time series of hydrological records. Only fundamental ideas, rather than details of individual tests, are presented here. Readers who are interested in a more detailed exposition may wish to consult a more extensive report (Kundzewicz & Robson, 2000), which is available free of charge upon request from the World Meteorological Organization, Hydrology and Water Resources Department, 7 bis av. de la Paix, Case Postale no. 2300, CH-1211 Geneva, Switzerland.

Acknowledgements The present article builds upon the material contained in the report edited by the authors and published under the auspices of the World Climate Data and Monitoring Programme (Kundzewicz & Robson, 2000). It is a contribution to a project entitled: "Analyzing Long Time Series of Hydrological Data and Indices with Respect to Climate Variability and Change" of the World Climate Programme—Water (WCP-Water Project A.2) coordinated by WMO and UNESCO. Support to ZWK from the Scientific Research Committee, Republic of Poland, under research grant no. 6 P 04 E 05 520 is gratefully appreciated.

REFERENCES

Bickel, P. M. & Van Zwet, W. R. (1978) Asymptotic expansion for the power of distribution-free tests in the two-sample problem. *Ann. Statist.* **6**, 987–1004 (corr. 1170–1171).

Bloomfield, P. & Steiger, W. L. (1983) *Least Absolute Deviations: Theory, Applications, and Algorithms*. Birkhauser, Boston, Massachusetts, USA.

Buishand, T. A. (1982) Some methods for testing the homogeneity of rainfall records. *J. Hydrol.* **58**, 11–27.

Buishand, T. A. (1984) Tests for detecting a shift in the mean of hydrological time series. *J. Hydrol*. **73**, 51–69.

Chiew, F. H. S. & McMahon, T. A. (1993) Detection of trend or change in annual flow of Australian rivers. *Int. J. Climatol.* **13**, 643–653.

Cleveland, W. S. (1994) *The Elements of Graphing Data*. Hobart Press, New Jersey, USA.

Davidson, A. C. & Hinkley, D. V. (1997) *Bootstrap Methods and Their Application.* Cambridge University Press, Cambridge, UK.

Efron, B. & Tibshirani, R. J. (1998) *An Introduction to the Bootstrap.* Chapman & Hall, London, UK.

Faulkner, D. S. & Jones, D. A. (1999) The FORGEX method of rainfall growth estimation. III: Examples and confidence intervals. *Hydrol. Earth System Sci.* **3**(2), 205–212.

Good, P. (1993) *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*. Springer-Verlag, Berlin, Germany.

Grubb, H. & Robson, A. (2000) Exploratory/visual analysis. In: *Detecting Trend and Other Changes in Hydrological Data* (ed. by Z. W. Kundzewicz & A. Robson), 19–47. World Climate Programme—Water, World Climate Data and Monitoring Programme, WCDMP-45, WMO/TD no. 1013. World Meteorological Organization, Geneva, Switzerland.

Helsel, D. R. & Hirsch, R. M. (1992) *Statistical Methods in Water Resources*. Studies in Environmental Science 49. Elsevier, Amsterdam, The Netherlands.

Hirsch, R. M. & Slack, J. R. (1984) A non parametric test for seasonal data with seasonal dependence. *Water Resour. Res.* **20**, 727–732.

Hirsch, R. M., Slack, J. R. & Smith R. A. (1982) Techniques of trend analysis for monthly water quality data. *Water Resour. Res*. **18**, 107–121.

Kundzewicz, Z. W. & Robson, A. (eds) (2000) *Detecting Trend and Other Changes in Hydrological Data*. World Climate Programme—Water, World Climate Programme Data and Monitoring, WCDMP-45, WMO/TD no. 1013. World Meteorological Organization, Geneva, Switzerland.

McGilchrist, C. A. & Woodyer. K. D. (1975) Note on a distribution-free CUSUM technique. *Technometrics* **17**(3), 321–325.

Pettitt, A. N. (1979) A non-parametric approach to the change point problem. *Appl. Statist.* **28**, 126–135.

Pilon, P. (2000) Criteria for the selection of stations in climate change detection network. In: *Detecting Trend and Other Changes in Hydrological Data* (ed. by Z. W. Kundzewicz & A. Robson), 121–131. World Climate Programme—Water, World Climate Data and Monitoring Programme, WCDMP-45, WMO/TD no. 1013, World Meteorological Organization, Geneva, Switzerland.

Radziejewski, M., Bardossy, A. & Kundzewicz, Z. W. (2000) Detection of change in river flow using phase randomization. *Hydrol. Sci. J.* **45**(4), 547–558.

Robson, A. J., Jones, T. A. & Reed, D. W. (1998) A study of national trend and variation in UK floods. *Int. J. Climatol.* **18**, 165–182.

Rousseeuw, P. J. & Leroy, A. M. (1987) *Robust Regression and Outlier Detection*. Wiley, New York, USA.

Siegel, S. & Castellan, N. J. (1988) *Non-parametric Statistics for the Behavioural Sciences*. McGraw-Hill, New York, USA.

Sneyers, R. (1975) *Sur l'analyse statistique des series d'observations*. WMO Technical Note no. 143. World Meteorological Organization, Geneva, Switzerland.

Sprent, P. (1989) *Applied Non-parametric Statistical Methods*. Chapman & Hall, London, UK

Tufte, E. R. (1983) *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, Connecticut, USA.

WMO (World Meteorological Organisation) (1988) *Analysing Long Time Series of Hydrological Data with Respect to Climate Variability and Change*. WCAP-3, WMO/TD no. 224, WMO, Geneva, Switzerland.

Worsley, K. J. (1979) On the likelihood ratio test for a shift in location of normal populations. *J. Am. Statist. Assoc.* **74**, 365–376.

Received 30 December 2002; accepted 4 November 2003